How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering

Year: 2020 | Citations: 502 | Authors: Zhengbao Jiang, J. Araki, Haibo Ding, Graham Neubig

Abstract

Abstract Recent works have shown that language models (LM) capture different types of knowledge regarding facts or common sense. However, because no model is perfect, they still fail to provide appropriate answers in many cases. In this paper, we ask the question, “How can we know when language models know, with confidence, the answer to a particular query?” We examine this question from the point of view of calibration, the property of a probabilistic model’s predicted probabilities actually being well correlated with the probabilities of correctness. We examine three strong generative models—T5, BART, and GPT-2—and study whether their probabilities on QA tasks are well calibrated, finding the answer is a relatively emphatic no. We then examine methods to calibrate such models to make their confidence scores correlate better with the likelihood of correctness through fine-tuning, post-hoc probability modification, or adjustment of the predicted outputs or inputs. Experiments on a diverse range of datasets demonstrate the effectiveness of our methods. We also perform analysis to study the strengths and limitations of these methods, shedding light on further improvements that may be made in methods for calibrating LMs. We have released the code at https://github.com/jzbjyb/lm-calibration.